通过病历首页收集肿瘤发病资料的数据清洗及效果评估

Data Cleaning and Effect Evaluation of Cancer Incidence Collected from First Page of Medical Recard // WU Fei, LIN Guo-zhen,ZHAN Jin-xin

吴　菲[1,2],林国桢[1],张晋昕[2]

(1.广州市疾病预防控制中心,广东 广州 510080;2.中山大学公共卫生学院,广东 广州 510080)

摘　要:从各级各类医疗机构的住院病案系统中调取病历首页信息,以发现新发肿瘤病例是肿瘤登记工作值得借鉴的方法。全文就广州市 2004~2009 年通过医院电子病历首页收集肿瘤发病资料的流程、数据清洗及其效果作一介绍。

关键词:肿瘤登记;发病资料;病历首页

中图分类号:R73-31　　**文献标识码**:A　　**文章编号**:1004-0242(2012)07-0507-03

肿瘤登记是系统地收集肿瘤信息的有效方式,是提供肿瘤病人整个病程资料及治疗效果的关键[1]。中国的肿瘤登记工作虽然取得了不少成绩,但登记覆盖人群面少,资料的完整性、准确性有待进一步提高[2]。从具有肿瘤诊断能力或收治肿瘤病人的各级各类医疗机构的住院病案系统中调取病历首页的信息然后对其进行整理出新发病例,是一项容易实施、病例收集较齐全的方法,但其非常复杂且工作量巨大,包括户口的确定、删重及利用死亡信息补充发病信息等。2004~2009 年广州市肿瘤登记中主要的新发病例均来自于医院的电子病历首页,现以 Excel 数据格式为例介绍数据清洗及其效果。

1　资料的整理方案

1.1　户口地址的确定

户口地址的确定是为了防止多报,即把登记范围以外的肿瘤病人列入统计资料内,从而高估登记区肿瘤发病率,户口地址确定的方案见图 1。

1.1.1　户口地址填写为非广州

户口地址填写为非广州的这类病例,可以确定户口为非广州。具体操作如下:对"户口地址"变量排序。注意要扩展到该病例的所有变量区域,防止只对该变量区域排序,在操作中要让首行变量处于筛选状态,然后在该变量的下拉菜单中选择"自定义排序"对其排序。若在操作中,不慎因未选到扩展区域,使得某些变量的信息与病人不对应,可用 VLOOKUP 函数,从原始库中批量找回相应的信息,公式为"=VLOOKUP(I:I,′1314.csv′! $B$1:$B$3000,1,0)"。

在"户口地址"变量旁插入一列,用于标注,方便最后的统一删除,如标注入选对象为"1"及删除的对象为"0"。由于已排序,所以在选择过程中可提高效率;有些地址填写不详细,如没有具体到区,要补充完整,同名的街道可用"替换"的功能,例如:晓港位于海珠区,把"晓港"替换成"海珠区晓港",但是对于

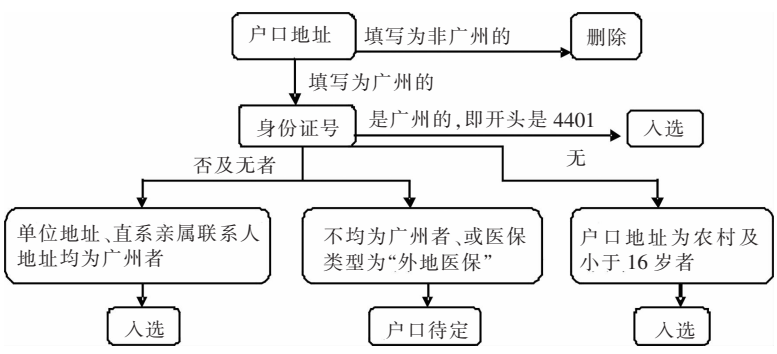

图 1　户籍确定技术路线图

收稿日期:2011-12-02

E-mail: feierwu@126.com

多个区有同名的街道,不能轻易替换,对于不确定的地址需进行电话随访确定。最后筛选出“0”进行统一删除。

1.1.2 户口地址填写为广州

户口地址填写为广州的病例中筛选出户口为非广州的病例。具体操作如下(在操作之前把第一步中插入的用于标注那列的内容清空):以“身份证号”变量排序,筛选出开头是 4401 的身份证号,并把这类病例入选,并在备注上标记为“1”。可用“筛选”选项中的“开头是”。用身份证号确定户口时,常常会发现有些身份证号显示成“3.30381E+17”的格式,所以要对单元格定义,即把单元格设定成“特殊”选项中的“邮政编码”格式,(注意每完成一次筛选操作之后要改回“全选”状态)。

身份证号非广州及无身份证号者的户口确定。筛选出需要操作的对象,即选择备注中的“空白”选项。籍贯、单位地址、直系亲属联系人(如夫妻、父子、母子)地址均为广州者入选为“1”,否则为“户口待定”,同时可结合“医保类型”协助判断。该操作中可以将不需要的变量“隐藏”起来,方便查看。

对于无身份证号者,而地址填写为农村的农村户籍,或年龄<16 岁者可直接入选为“1”,这类对象是外地户口的可能性很小。具体操作如下:①先选择“身份证号”变量中的“不详”、“无”、“空白”等选项筛选出无身份证号者,再选择“户口地址”中同时包含“镇”和“村”的农村户口地址对象标注为“1”。注意有些以“村”命名的地方不一定是农村,如越秀区的“梅花村”。②筛选出<16 岁者入选为“1”。对于没有填写年龄而有出生日期和身份证号者,可通过身份证号推导出生日期,然后用出生日期和入院日期计算发病时的年龄。先把出生日期的单元格确定为日期格式,再用公式:“=TEXT (MID (A1,7,4)&"-"&MID (A1,11,2)&"-"&MID (A1,13,2),"YYYY-MM-DD"”生成出生日期,然后填充全列,其中 A1 为身份证号。然后,年龄的格式先确定为常规,所有的日期单元格格式转换成日期,用公式=DATEDIF(B1,C1,“Y”)计算出年龄,然后填充全列其中 B1 为出生日期,C1 为入院日期。如果日期中有具体时间的,需去除时间才能用上述公式,即在日期旁插入一列,单元格格式变为常规,用公式“=A1*1”或“=left(A1,10)”,然后填充全列,其中 A1 为含有时间的“日期”,注意要把单元格格式转变回日期格式,生成的新的日期选择性粘贴,选择“数值”。

1.1.3 对“户口待定”的病例进行电话随访

这个步骤应在“删重”完成后进行,以减少重复的随访工作量。电话随访前,先与肿瘤登记系统进行比对,因为广州市肿瘤登记系统已经开展部分随访工作,这样可减少电话随访的工作量。用 VLOOKUP 函数对户口待定的病例与肿瘤系统导出的数据库进行比对。同时,对每一个调查员进行严格的培训,每个随访不成功的病例都应保证随访 3 次后才放弃。对于因缺少联系电话或电话号码错误而无法确定的病例,返回医院让医院核实。

1.2 删 重

删重包括对当年所有医院的肿瘤病例的重复病例进行删除,并且要甄别出当年肿瘤病例,即与历史发病库里的记录进行比对,判断病例是否为当年的新发病例。

确定同一病例的原则:主要根据身份证号、出生日期、性别、户口地址、联系人姓名、联系人地址、疾病诊断,由主到次综合判断。对于同名但不能确定为同一病例的在备注上标明“查重待定”,以备进一步电话随访;一旦确定为同一病例,逐条记录逐个变量比较合并。

合并原则:①从医院病案首页导出的所有变量种类和排序(1=病案号,2=姓名,3=性别,4=年龄,5=出生日期,6=身份,7=民族,8=职业,9=婚姻情况,10=入院次数,11=入院日期,12=确诊日期,13=出院日期,14=疾病名称,15=ICD-10 编码,16=M 编码,17=病理诊断,18=是否手术,19=是否尸检,20=转归,21=是否随诊,22=随诊期限,23=住院医师,24=主治医师,25=主任医师,26=门诊医师,27=籍贯,28=户口地址,29=户口邮编,30=单位名称,31=单位地址,32=单位邮编,33=单位电话,34=联系人,35=关系,36=联系人地址,37=联系人电话,38=放射费,39=报告医院,40=备注),首先保留最早的入院日期和确诊日期,保留最晚的出院日期。②疾病诊断和病理诊断:保留描述最为详细或大医院的疾病诊断和病理诊断。保留原发肿瘤,如为多原发肿瘤,保留多条记录(即这几条记录其他信息一致,只有疾病诊断、病理诊断及编码不一致),若只有继发肿瘤,保留其中一个诊断。保留诊断名称时需要把相应的 ICD-10 编码、病理诊断及 ICD-O-3 编码一起保留。多原发的每条记录在备注 2 上标注“2”,只有继发及转移

的标注“3”。③转归：保留最后一次出院时的情况。当最后记录为“死亡”时，同时将“出院日期”更改为最后记录，因为死亡时间以该次住院的出院日期为准。④出生日期：几条记录不一时，按照身份证号上日期。⑤职业：选择有具体工种的记录。⑥户口地址及户口邮编、联系人及联系方法：选择保留最详细的最新的信息。⑦其他变量均保留最全的信息：如是否手术，只要曾经有手术，保留“是”；只要有医院进行了随访的保留随访记录入院次数：最多的一次。⑧如果同一病人在往年发病了，但是有新的原发疾病要保留。

删重的具体操作：①首先对“姓名”进行排序，然后在“姓名”旁插入一列，单元格格式改为常规，在第一个名字旁的单元格输入公式“=IF(D2=D1,0,1)”，填充全列，此时于前一个病人同名的病例显示“0”，不同名的病例显示“1”，再在旁边插入一列，把运用公式得到的结果的“值”保留，即“粘贴值”，同时把公式列删除，最后，把“0”的记录筛选出，标为蓝色，以便删重时更直观。②把往年数据库的字体标红后导入，只需保留往年资料中用于确定是否为同一病例的基本信息，减少信息量使 Excel 运行更顺畅，重新运用查重公式。③只在当年数据中出现同名病例，即蓝色和黑色的同名病例，按照合并原则进行删重合并；往年数据中出现同名病例，即红色字体显示“0”时，与当年数据比对，只要确定是同一病例，就可把当年病例中的“1”改为“0”以备最后的删除，无需进行删重合并。不能确定是否为同一病例的同名病例，在插入列中改为“4”。④最后，将“0”筛选出并删除，然后再把往年数据库，即红色字体筛选出并删除。

2 质量控制和方案评价

为评价户籍确定方法的可靠性，我们从整理确定后的病例中（抽查的总体不包括电话随访确定的病例），通过简单随机抽样的方法分别抽查 200 例入选的病例和 200 例删除的病例，随机数字由 SAS 软件生成，对这 400 例病例通过电话访问的方式进行核查，结果显示，删除的抽样样本中，错误病例数为 0，即假阴性率为 0；入选的抽样样本中有 7 例病例信息确定有误，经查询，这 7 例的信息错误是人为的失误。该方案的特异性为 96.5%，灵敏度为 100%，约登指数为 96.5%，说明这种综合确定方案的灵敏度、特异度和真实性都很高。该方案的 Kappa 值为 0.965，说明这种确定方案同金标准的结果一致性很高。

3 讨 论

提高病例资料的质量关键在于提高病例报告卡和随访卡填写完整性和准确性，同时最大限度降低肿瘤漏报率。临床医生对信息完整性和准确性的保证起着关键的作用，需要临床医生引起高度重视[3]。目前上报全国肿瘤登记中心数据的登记处共 50 个，其中资料质量符合基本要求的有 39 个，覆盖人口近 7 千万，达全国人口的 6%。有 20 个省开展了肿瘤登记工作，登记质量也较以前有明显提高[4]。目前，大连市、北京市肿瘤登记模式及网络信息系统建设不断完善，但是仅广泛应用于全市二级及以上医院的住院患者和全市因肿瘤死亡的死亡病例，漏报的门诊病例依靠后期住院治疗和提及肿瘤史的死亡病例补充（死补活）[5,6]。除北京、上海、大连等大城市可以直接用肿瘤登记系统中导出的数据，全国大部分地区仍然是通过病案首页导出的数据来整理新发病例资料。

未来户籍的确定可以利用身份证、医保证等，但病例重复肯定长期存在，而且随着病人生存时间延长和流动性增加，重复看病更多。肿瘤登记系统中的数据整理，即户籍确定和删重的原则与这个方法也是一样的。广州市肿瘤登记工作中的数据整理的经验，可以提供给相关部门借鉴。

参考文献：

[1] Middleton RJ, Gavin AT, Reid JS, et al. Accuracy of hospital discharge data for cancer registration and epidemiological research in Northern Ireland [J]. Cancer Causes Control, 2000, 11(10): 899–905.

[2] 李伟栋, 张晋昕. 肿瘤登记报告卡的质量控制 [J]. 中国肿瘤, 2010, 19(12): 782–785.

[3] 郑莹, 沈玉珍, 李德録, 等. 肿瘤报告质量的规范管理[J]. 中国肿瘤, 2002, 11(6): 6–7.

[4] 陈万青. 中国肿瘤登记系统的建立与完善[J]. 中国肿瘤, 2011, 20(1): 7–9.

[5] 张莉梅. 大连市肿瘤登记模式完善与网络信息系统建设[J]. 中国肿瘤, 2009, 18(5): 356–358.

[6] 王宁, 祝伟星, 邢秀梅. 北京市肿瘤登记信息系统建设和完善[J]. 中国肿瘤, 2010, 19(3): 150–154.